

作者簡介

里克是德國弗里堡大學（University of Fribourg）的機器學習教授，並擔任 AI 資料科學實驗室（AI for Data-Oriented Science）首席研究員。他研究如何透過幾何學和拓樸學的視角，改善機器學習和 AI 所需要的堅實基礎。

機器學習正在成為我們這個時代的變革性技術：我們當中的許多人都曾與 ChatGPT 等模型互動過，醫療保健等應用領域幾乎每天都在宣布新的突破，而且一些數學家已經使用大型語言模型作為證明助手，正在開闢一條將這些工具整合到科學研究中的新途徑 [Ghr25, Won24]。

本文是專門為了對機器學習了解有限的數學家讀者所撰寫的，目的在消除一些關於機器學習領域的誤解，使其更具開放性和包容性。誠然，從表面上看，機器學習似乎是一個單質化的領域，但實際上，它就像數學本身般多元豐富。雖然該領域的主要發展動力源自於工程技術的進步，例如建立更龐大精良的模型，但數學和數學家在此領域也擁有廣闊的發展空間。本文將以拓樸學為例，闡述數學（超出線性代數和統計學這些機器學習的經典驅動力之外）如何豐富該領域。拓樸學近期開始為機器學習的基礎理論提供新的視角：點集拓樸利用鄰域等概念，可以將現有演算法從圖擴展到胞腔複形（cell complex）[HIZ20]；代數拓樸利用同調等有效不變量，可以改善 3D 形狀重建模型的性能 [WAM+22]。最後，微分拓樸為研究資料的光滑性質提供了工具，從而產生了分析幾何單體複形（simplicial complex）的有效方法 [MHR24]。這些（以及更多）研究方向正在新興的拓樸深度學習領域中找到歸宿 [PBB+24]。

1. 何謂機器學習？

在深入探討具體例子之前，我們先退一步，介紹一下機器學習領域。機器學習通常被理解為運用演算法解決任務的藝術和科學，但是加上了限制演算法應該——在某種意義上（下文將闡明）——能夠因應手頭上的資料。為了闡明這個非正式的定義，我們需要引入一些術語，首先是特徵（feature）的概念，它指的是從「原始」（raw）輸入資料中提取的量。例如，給定一組化學結構，碳原子數可以作為一個特徵。加入其他特徵——例如氧原子數、環狀結構或芳香鍵——就能得到化學結構的高維度表示，即特徵向量（feature vector）。為了方便起見，我們假設該特徵向量存在於某個 $\mathbb{R}^d$ 空間中。

因此，從宏觀層面來看，機器學習演算法可以被視為形式為 $f: \mathbb{R}^d \rightarrow \mathcal{D}$ 的函數，其中 $\mathcal{D}$ 表示某個感興趣的領域。當 $\mathcal{D} = \mathbb{R}$ 時，我們稱之為迴歸任務（regression task）^①；而當 $\mathcal{D}$ 是集合時，我們處理的是分類任務（classification task）。例如，假設我們對化學結構進行分類感興趣。定義域 $\mathcal{D}$ 可以是由「有毒」和「無害」這兩個標籤組成。由

① 註：同樣適用於預測多個值的情況。

作者鳴謝：作者衷心感謝羅爾（Ernst Röell）、西蒙斯（Emily Simons）以及匿名審稿者提出的寶貴意見，這些意見大幅提升了本文的品質。

本研究由瑞士聯邦教育、研究與創新國務秘書處（Swiss State Secretariat for Education, Research, and Innovation）所資助。